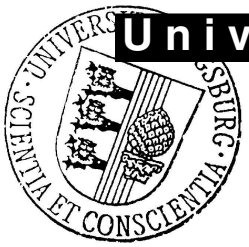

Universität Augsburg

Institut für Mathematik

Antony Unwin, Martin Theus, Wolfgang Härdle

Exploratory Graphics of a Financial Dataset

Preprint Nr. 031/2007 — 13. Juli 2007

Institut für Mathematik, Universitätsstraße, D-86 135 Augsburg http://www.math.uni-augsburg.de/

Impressum:

Herausgeber:

Institut für Mathematik
Universität Augsburg
86135 Augsburg
http://www.math.uni-augsburg.de/forschung/preprint/

ViSdP:

Antony Unwin
Institut für Mathematik
Universität Augsburg
86135 Augsburg

Exploratory Graphics of a Financial Dataset

Antony Unwin[1], Martin Theus[2] and Wolfgang Härdle[3]

[1] Augsburg University, Germany unwin@math.uni-augsburg.de
[2] Augsburg University, Germany martin.theus@math.uni-augsburg.de
[3] Humboldt-Universität zu Berlin, Germany haerdle@wiwi.hu-berlin.de

1 Introduction

The first stages of any data analysis are to get to know the aims of the study and to get to know the data. In this study the main goal is to predict a company's chances of going bankrupt based on its recent financial returns. In another chapter of the Handbook, some sophisticated prediction models based on support vector machines are discussed for a similar dataset. Here, visualization methods are used to explore the large dataset of American company accounts that was made available for predicting bankruptcy, to get to know the data, and to assess the quality of the dataset. This is an initial exploratory analysis not using any expert accounting knowledge.

Exploratory Data Analysis (EDA) has been a well-known term in Statistics since Tukey's historical book (Tukey; 1977). While everyone acknowledges the importance of EDA, little else has been written about it, and modern methods, including interactive graphics (Unwin; 1999), are not as much applied in practice as they might be. Interactive graphics were used extensively in the exploratory work for this chapter. The dataset is by no means particularly big but does contain more than 80,000 cases. Ways of graphically displaying large datasets are discussed in detail in Unwin, Theus and Hofmann (2006).

In considering graphic displays, it is necessary to distinguish between presentation and exploratory graphics. Graphics for displaying single variables or pairs of variables are often used for presenting distributions of data or for presenting results. Care must be taken with scales, with aspect ratios, with legends, and with every graphic property that may affect the success of the display in conveying information to others. Graphics for exploration are very different. They are more likely to be multivariate and there is no need to be particularly concerned about the aesthetic features of the graphic; the important thing is that they give a clear representation of the data. Presentation graphics are drawn to be inspected by many people (possibly millions of people, if they are used on television) and can be long-lived. Playfair's plots (Playfair; 2005) of English trade data, for example, are over 200 years old but

are still informative. Exploratory graphics are drawn for one or two people and may be very short-lived. A data analyst may examine a large number of graphics, before finding one that reveals information, and, having found that information, may decide that another kind of display is better for presenting it than the display or displays that led to its discovery.

The graphics shown in this chapter are a subset of those used in the study. They are not supposed to be "pretty", they have been drawn to do a job. Detailed scales have not been included, as it is always the distributional form that is of primary interest. Exploratory analyses are subjective and each analyst may choose different graphics and different combinations of them to uncover information from data. The only thing that can be said with certainty is that analysts who do not use graphics to get to know their data will have a poor understanding of the properties of the data they are working with. If nothing else, graphics are extremely useful for assessing the quality of data and for cleaning data.

2 Description of the Data

There are 82,626 records in the dataset. Each record contains financial information for a company for one year. Table 1 gives a list of the variables in the dataset.

Variable	Description
Cash.TA	Cash/Total Assets
Inv.TA	Inventories/Total Assets
CA.TA	Current Assets/Total Assets
Kap.TA	Property, Plant and Equipment/Total Assets
Intg.TA	Intangibles/Total Assets
LogTA	log(Total Assets)
CL.TA	Current Liabilities/Total Assets
TL.TA	Total Liabilities/Total Assets
Eq.TA	Equity/Total Assets
S.TA	Sales/Total Assets
Ebit.TA	EBIT/Total Assets
Ebit.Int	EBIT/Interest Payments
NI.TA	Net Income/Total Assets
CA.CL.TA	(Current Assets - Current Liabilities)/Total Assets
BANKR	status (Bankrupt or OK)
Year	year of accounts
State	where the company was registered
NAICS	North American Industry Classification System

Table 1. *Variables in the Bankruptcy dataset*

For each company there are 13 ratios, one size variable (log transform of assets) and a binary variable, which records whether the company went bankrupt within three years of the financial returns or not. There is also information on the State the company is registered in, on the industry sector and on the year of the accounts. The term "bankruptcy" includes Chapter 11 reorganization as well as liquidation under Chapter 7 of the US Bankruptcy Code. Financial ratios are commonly used so that data are comparable across years. Sometimes it is helpful to look at the raw data as well, particularly if there are data cleaning issues. An unusual value, or cluster of unusual values, in *Total Assets*, would affect all twelve ratios dependent on this variable. There were no missing values. Some of the ratio variables can be grouped into categories: Profit Measures (*Ebit.TA*, *NI.TA*); Leverage ratios (*Kap.TA*, *TL.TA*, and *Eq.TA*); Liquidity ratios (*Cash.TA*, *CA.TA* and *CA.CL.TA*); and Activity/Turnover ratios (*Inv.TA*, *S.TA*, and *Ebit.Int*).

To be able to generalise results from statistical models, the dataset analysed has to be a random sample from the population under study. For the bankruptcy dataset, there is a very large number of cases, but it is not clear how far they can be regarded as a random sample from anything. Apart from anything else, it is not at all clear that they can be considered to be homogeneous. Most companies are rather small and a few are very large. Can the same financial ratios really be used for companies that are so different in scale? This is the kind of question that exploratory analysis can help answer by looking at the distributions of values of data for the different groups.

The assumption is that results from this dataset can be used on datasets collected in similar fashion in the future.

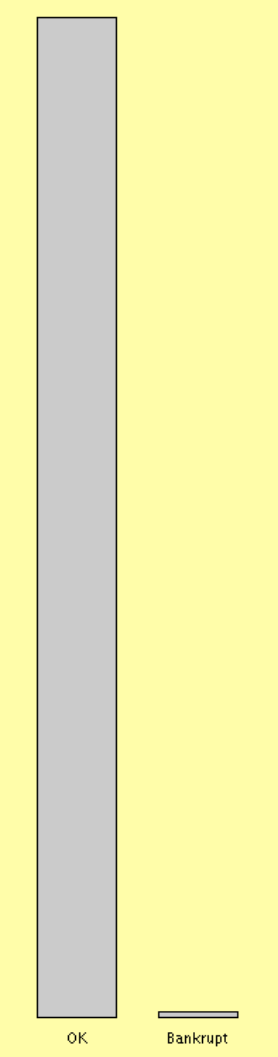

Fig. 1. A bankruptcy barchart. 506 of the 82626 records refer to bankruptcy.

3 First Graphics

Figure 1 is a barchart for the bankruptcy variable. Only a small proportion of companies actually went bankrupt, which is a good thing for the companies, but makes any statistical analysis more difficult.

The displays in Figure 2 and Figure 3 show that the companies are fairly equally distributed across the regions but that some of the data are surprisingly old with a few cases prior to the mid 1960's. In the 1950's there is one record per year and linking to the geographic data shows it was not always the same company, as might have been suspected.

The geographic information was originally provided by State, so there were a large number of small counts and only a few big ones. Grouping by region gives a good overview, though other groupings (e.g., by population) could also be tried. The regional classification used here is one from the FBI. Selecting the foreign group, a little more than 10% of the cases, and linking to a spinogram (Hofmann and Theus; 2006) of years, (see Figure 4), shows that the percentage of foreign registered companies has increased over the years. Querying shows this to be from about 4% in the early 1970s to 15% in 2002. In the most recent year, 2003, the rate falls to just under 11%. It is expensive for foreign firms to be listed on the US exchanges and opinion has changed as to what benefits it brings. The Sarbanes-Oxley Act has also made it less attractive to be listed in the US.

Information was also available on industry sector in two different ways, with one classification of 426 categories by name and numerical coding and the NAICS classification by number with 925 categories. Both are too detailed for graphical analysis and a hierarchical grouping similar to the spatial grouping of the States can be tried. The six-digit NAICS codes can be aggregated by their first two digits and then further grouped by sector to give Figure 5. The manufacturing sector clearly dominates. To check for associations between the two classifications a crude scatterplot approach was tried. A fairly random spread of points was obtained but no particular pattern.

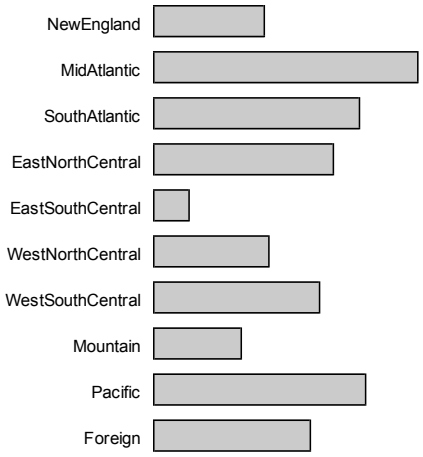

Fig. 2. A barchart of the numbers of records by US region.

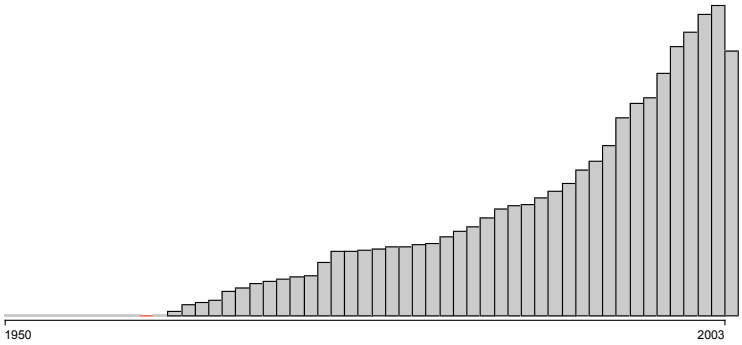

Fig. 3. A histogram of the numbers of records per year.

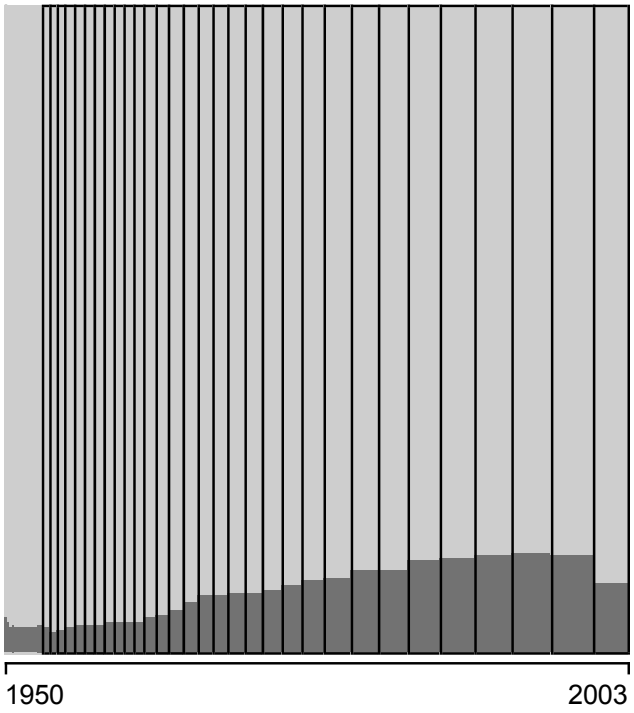

Fig. 4. A spinogram of the numbers of records per year with foreign registered companies selected. The width of a bar in a spinogram is proportional to the height of the corresponding bar in the original histogram.

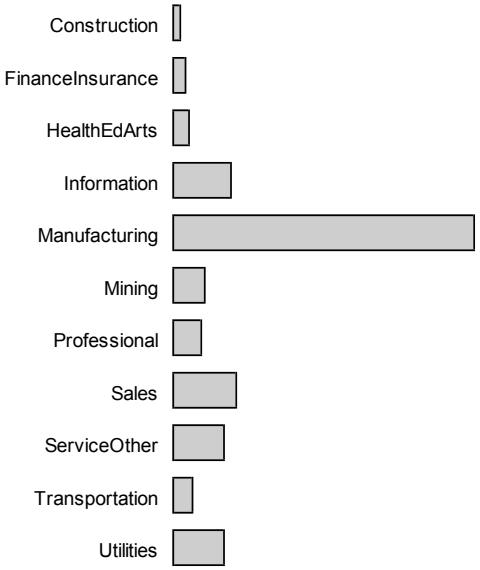

Fig. 5. A barchart of the numbers of records by NAICS groups.

Histograms might be drawn for the continuous variables, but they take up quite a lot of room, so boxplots are more efficient for displaying many variables at once. Since *logTA*, the log of total assets, is a different kind of variable from the others and important in its own right as well (because it groups the companies by size), a histogram has been drawn just for it (see Figure 6). This shows a roughly symmetric distribution with a few very low values. Selecting the foreign-registered companies again and linking to a spinogram of *logTA* reveals that the percentage of these companies rose steadily for the biggest 25% of the companies, from 7% up to more than 50%.

A set of parallel boxplots for the ratio variables is shown in Figure 7. The boxplots reveal that several of the ratios are highly skewed and this may affect whether they can be of much help in discriminating between the companies which went bankrupt and those which did not. It is possible that many of these outliers are errors of some kind and it may be that most of the outliers on individual variables are actually outliers on several variables.

4 Outliers

Outliers can be checked with a parallel coordinates display (Inselberg; 1999), as in Figure 8, where the eight ratios with highly skew distributions have been plotted and seven of the worst outliers selected. It is easy to see that several

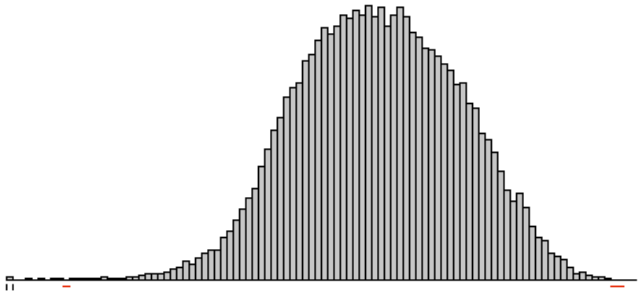

Fig. 6. A histogram of log(total assets), *logTA*, for the companies. The marks to the left under the axis are interactive controls for the anchorpoint and binwidth. The horizontal (red) marks record bins where the count is too small to be drawn.

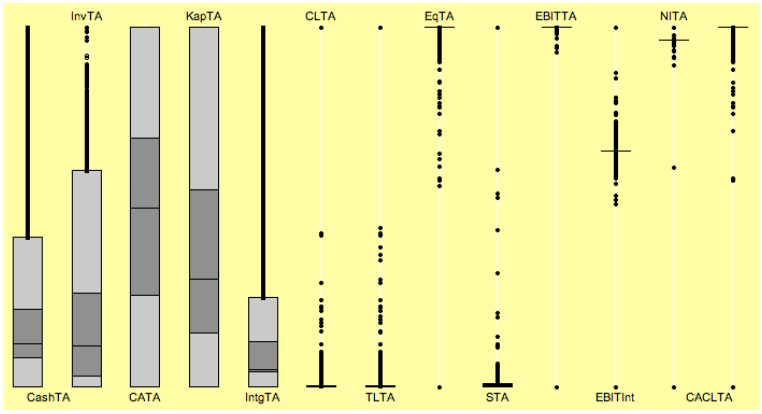

Fig. 7. Parallel boxplots of financial ratios. Each boxplot is individually scaled.

are outliers on more than one variable. It is also apparent that the ratio of equity to total assets *Eq.TA* is perfectly inversely correlated with the ratio of total liabilities to total assets *TL.TA*. This is a matter of definition with $TL.TA + Eq.TA = 1$. Although the equation looks innocuous it masks the fact that in this dataset *TL.TA* ranges from 0 to 5838. Not surprisingly, the high values of *TL.TA* only arise for low values of *Total Assets* as the L-shaped scatterplot on the left of Figure 9 shows. This plot is somewhat misleading. The zoomed version on the right reveals that there is more variability amongst the low values than the default plot suggests, though the most extreme values of *TL.TA* are still only for very low values of *Total Assets*. The bulk of the pattern suggests that very small companies have a broader range of possible liability ratio values than small companies. The low density region to the lower right implies that companies exceeding a certain size must have some liabilities. The distorting effect of the extreme values is demonstrated by the

fact that the zoomed plot, which is 10^{-5} the size of the default plot, contains 87% of the data. α-blending has been used to make the distributional structure more visible. α-blending weights the shading of each object by a specified fraction. The darkness of a pixel is that fraction times the number of objects overlapping the pixel or 1, whichever is smaller.

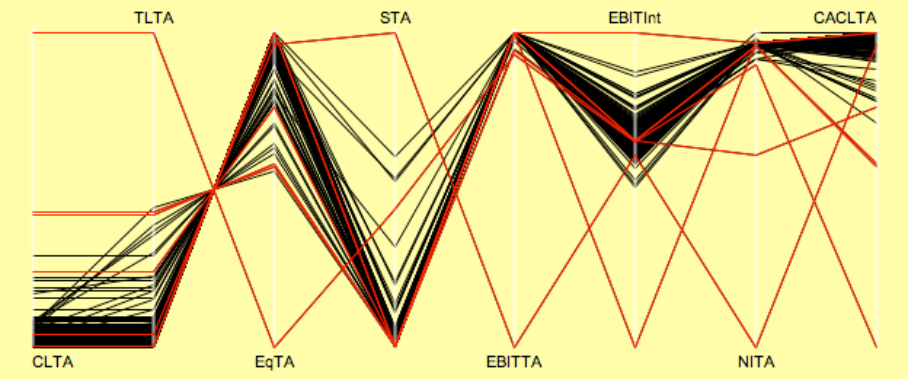

Fig. 8. Parallel coordinate plot of financial ratios with skew distributions. Seven outliers have been selected.

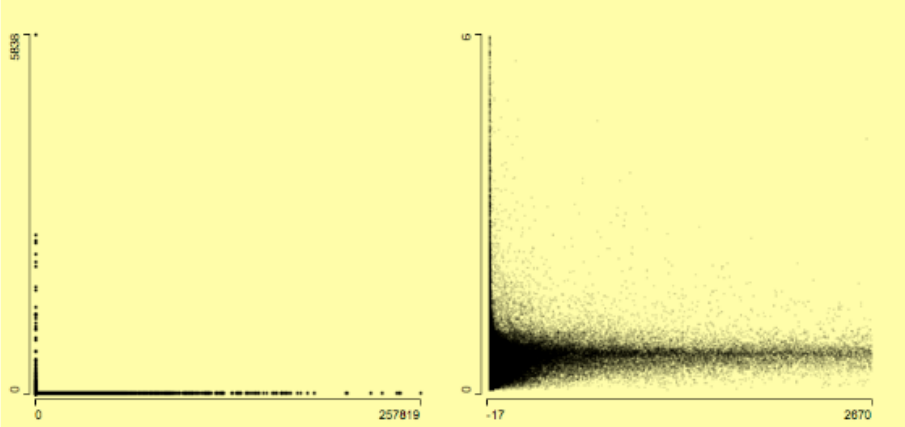

Fig. 9. Scatterplots of *TL.TA*, the ratio of total liabilities to total assets, plotted against *Total Assets*. On the left all the data, showing that all high values of liabilities are associated with low values of assets. On the right, a zoom of about 10^{-2} on the x-axis by 10^{-3} on the y-axis with some α-blending.

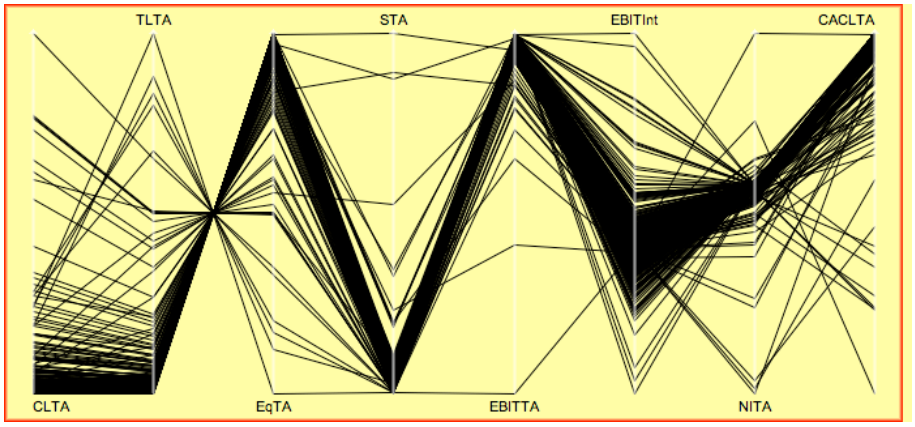

Fig. 10. Parallel coordinate plot of the financial ratios with skew distributions. The seven outliers selected in Figure 8 have been removed. The plot's (red) border is a sign that not all data are displayed.

Setting aside the seven worst outliers, Figure 8 rescales to Figure 10. Some of the variables become potentially more informative, the scales of some others are dominated by newly visible outliers.

Outliers can be dealt with in several ways. A transformation might help (but that is not always appropriate, in this case several ratios have some negative or zero values). The outliers could be trimmed or discarded (but that depends on what kind of outlier they are). With ratio variables, as here, the component variables of the ratios can be examined. Figure 11 shows a scatterplot of the variable *Sales* against *Total Assets* with the same seven outliers still highlighted. All turn out to be small companies, in terms both of *Sales* and of *Total Assets*, and small companies should probably be treated separately anyway. The scatterplot also reveals that there are some bivariate outliers. Querying shows that the six cases with high *Sales* but low *Total Assets* (the group upper left) are all the same big retail company over several successive years. The seven cases with low *Sales* but very high *Total Assets* (the group lower right) represent three companies from the information and communication sector. These two groups make up a tiny proportion of the dataset, even if they are all very large on either *Sales* or *Total Assets*.

For exploratory work it is distributional structure that is of interest, not precise values. The minima and maxima of the variables are used to determine the limits of the axes. The lower limit for *Sales* in Figure 11 shows that the lowest *Sales* values are actually negative and that there are plenty which are zero or almost zero as well. The negative values could be an accounting quirk, but should surely be discarded from the dataset, though there are only five of them. The low *Sales* values are also worth considering, what kind of companies are these and should they be kept in the dataset? They may be very new companies or companies on their last legs. There are 1412 companies

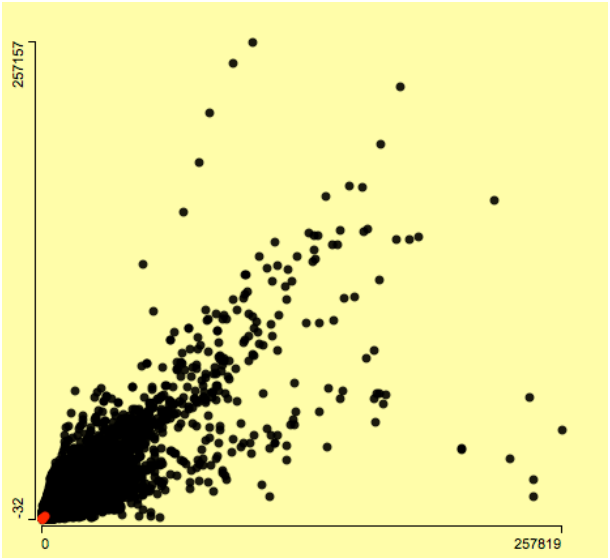

Fig. 11. Scatterplot of *Sales* v. *Total Assets* with the 7 outlying companies highlighted, the lighter (red) blob in the lower left corner.

with zero *Sales* and another 3674 with *Sales* more than zero but less than 1. These data could in principle have been obtained by zooming into a histogram for *Sales* and querying the cells, but for queries with precise boundaries it is quicker to calculate the appropriate frequency table. Graphics are better for more general, qualitative insights, while tables and statistical summaries are better for exact details.

The form of empirical distributions can be examined in many ways. When cases are of different sizes or weights, it can be revealing to look at weighted distributions. For instance, Figure 12 shows a histogram of *CA.TA*, the ratio of current assets to total assets, on the left and a histogram of the same variable weighted by *Total Assets* on the right. Companies with the highest current assets ratios clearly have low *Total Assets*.

Outliers and negative values are some of the data cleaning problems that can arise, there may be others as well.

Some statistical modelling approaches are little affected by individual gross errors and it may be that whether these cases are excluded, or adjusted, or just included in the analysis anyway, will not affect the model fit. Even when this is the case, it is useful to know what kinds of errors or unusual values can occur. It is also an opportunity to talk to people who know the context of the data well, and to get them to provide more background information. Analysing data blind without having any information is just reckless. It is surprising (and sometimes shocking) how much useful information is known about datasets that is not incorporated into analysis, be it information about

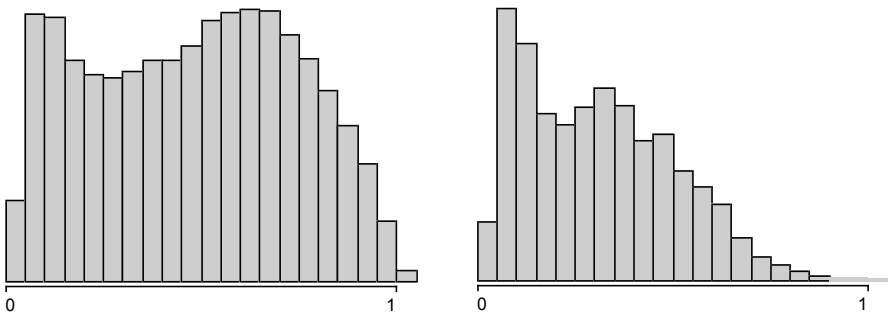

Fig. 12. A histogram of the current assets ratio on the left and a weighted histogram of the same variable, weighted by *Total Assets*, on the right.

the selection of variables collected, about the way the sample was chosen, or about the details of the data collection and data recording. Dataset owners may assume that analysts know this information but it is useful to check.

5 Scatterplots

There is often a temptation with large numbers of continuous variables to calculate correlation coefficients for all possible pairs of variables. The trouble is that correlations can be high because of outliers, or low because the association between the two variables is non-linear. Fourteen variables (the original ratios plus *logTA*) are too many to draw a scatterplot matrix, but it is interesting to look at the associations between subsets of the ratios. Scatterplots for two pairs of the financial variables *Cash.TA*, *Inv.TA*, *CA.TA* and *Kap.TA* are shown in Figure 13. The lower left triangular shapes show that the sum of the corresponding ratios is less than a limit, in these cases 1. (Lower right triangular shapes are obtained when the y variable is always lower in values than the x variable, as when the cash or inventory ratios are plotted against the current assets ratio.) The negative cash values that have already been discussed are easily seen to the left of the bulk of the data in the plot of cash and inventory ratios. In the plot of the current assets and property ratios there are a few companies which surprisingly lie above the bounding diagonal — more outlying cases to investigate. The correlation coefficients for the variables are given in Table 5. Some of them are quite high, they are certainly all significantly different from zero, but they hint little at the structure shown. The highest value in absolute terms is the correlation between current assets and property (-0.752). This deserves further investigation and Figure 14 shows the same scatterplots as in Figure 13, this time using the smallest pointsize and some α-blending instead of the defaults, providing a rough bivariate density estimate. It is now possible to see that for many of the companies, the sum of current assets and property assets is almost equal to *Total Assets*.

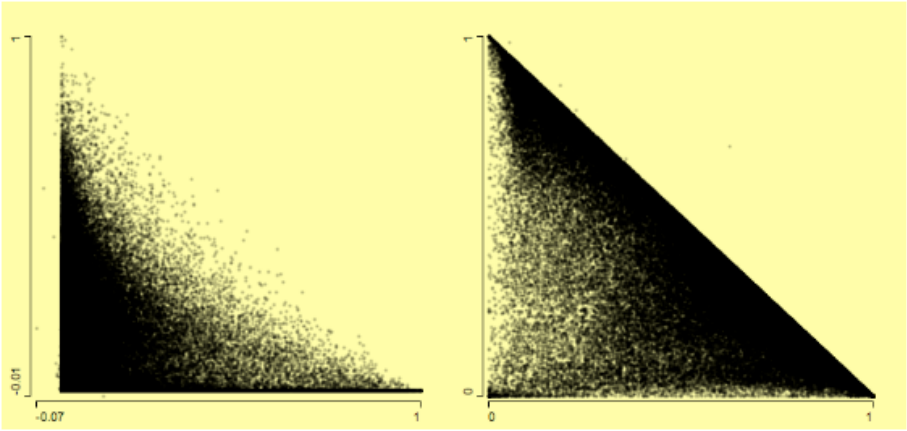

Fig. 13. Scatterplots of *Cash.TA* and *Inv.TA*, the ratios of cash and inventories to total assets, (left), and of *CA.TA* and *Kap.TA*, current assets and property to total assets (right).

	Cash.TA	Inv.TA	CA.TA	Kap.TA
Cash.TA	1.000			
Inv.TA	-0.224	1.000		
CA.TA	0.482	0.553	1.000	
Kap.TA	-0.378	-0.352	-0.752	1.000

Table 2. Correlations between the four ratio variables in Figure 13.

The scatterplots of the corresponding raw data variables provide another view of the associations between variables in the dataset, and more structure

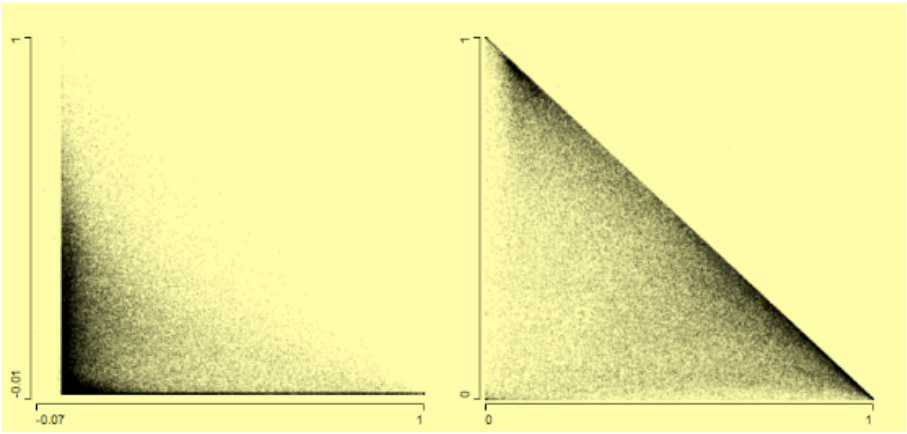

Fig. 14. The same scatterplots as in Figure 13 but with a smaller pointsize and α-blending to display the bivariate structures better.

can be seen than with scatterplots of the ratios. A few of the raw scatterplots exhibit a funnel structure, such as the plot of inventories against fixed assets and that of cash against current assets (Figure 15). Querying and linking can be used to identify specific sectors or outliers. The companies with high assets and low inventories are in the information and communication sector. Companies in the retail sector have higher inventories and lower assets. The at first sight unusual three points to the top left in the inventories/fixed assets plot are from investment banking and are consistent with other companies in that sector — except for being considerably bigger. The biggest companies on both variables in the left-hand plot are car manufacturers (higher inventories) and oil companies (higher assets).

Sometimes features stand out in a parallel coordinate plot and sometimes they are more visible in raw data scatterplots. As always, a range of different graphics should be explored.

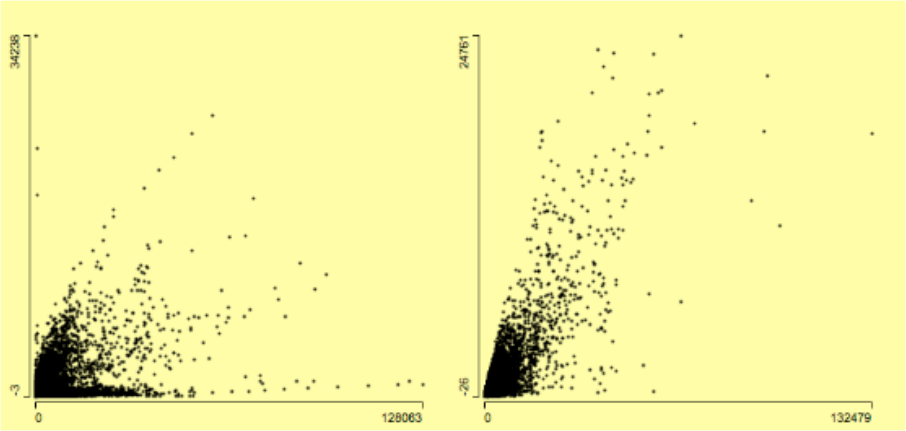

Fig. 15. Scatterplots of inventories against fixed assets (left) and of cash against current assets (right).

6 Mosaic Plots

Combinations of categorical variables may be displayed in one form or other of mosaic plots (Hofmann; 2000). For this dataset there are two difficulties with drawing these plots. Firstly, all but one of the categorical variables have very many categories (there are 54 states and 925 NAICS categories). Secondly, the one categorical variable that is binary (*BANKR*) has less than 1% of the cases in one category, so that highlighting is rarely visible. The first problem can be partly got round by grouping, combining states into regions and using

a less detailed form of the NAICS. The second problem could be solved by a special form of zooming.

Figure 16 shows a fluctuation diagram of the numbers of companies by industry sector and region. Classical mosaic plots try to make the most efficient use of the space available by making each cell as large as possible. This can make the plot difficult to interpret, especially when there are many cells. Fluctuation diagrams preserve a grid structure, so that comparisons by rows and columns are easier. The dominance of the manufacturing sector stands out in Figure 16, as well as individual details, such as the concentration of mining companies in WestSouthCentral.

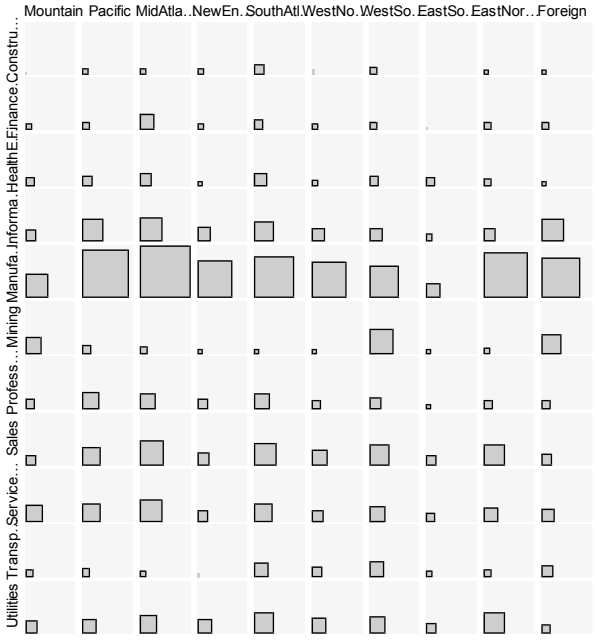

Fig. 16. Fluctuation diagram of industry sectors by regions.

All cases are treated equally in Figure 16. Of course, some companies are much bigger than others and Figure 17 shows the same cases weighted by *Total Assets*. The differences between the two figures are striking. Foreign registered companies in the manufacturing sector are much bigger than was apparent before. The regional distribution of companies in the professional sector changes a lot.

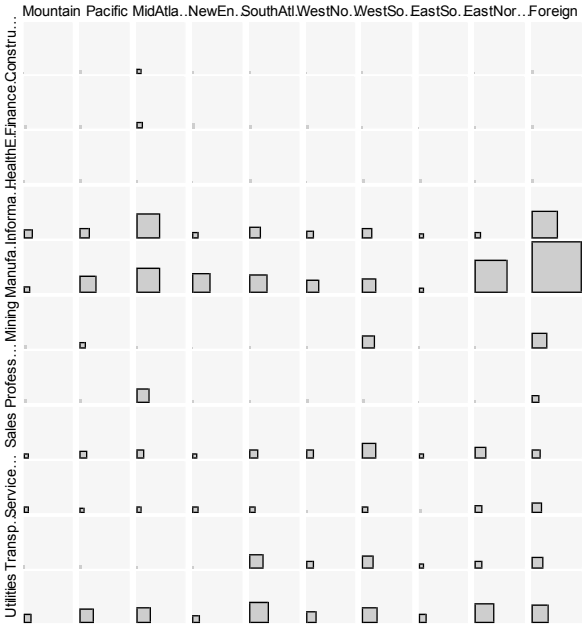

Fig. 17. Fluctuation diagram of sectors by regions weighted by *Total Assets*.

7 Initial Comparisons of Bankrupt Companies

The analysis up till now has considered all the data as one group and a number of outliers, distributional features, and specific properties have been identified. The main aim of the study remains the investigation of how the companies which went bankrupt differ from the rest. A first approach would be to look at the information available. Any company whose total liabilities exceed their total assets ($TL.TA > 1$) is likely to be in trouble. This variable is bizarrely skew and so difficult to visualize with standard plots. In this situation a simple table is better.

	Bankrupt	OK
TL.TA > 1	272	5856
	4.44%	95.6%
TL.TA ≤ 1	234	76264
	0.31%	99.7%

Table 3. Cases with more liabilities than assets and their bankruptcy status.

Clearly the variable *TL.TA* on its own is very informative. The choice of the boundary limit of 1 is determined by the context. Varying the limit interactively using a slider confirms that it is a good choice, in that both the difference in bankruptcy rates and the number of bankrupt companies are high.

To investigate the effects of more than one variable, two kinds of parallel plots can be used with the bankruptcy cases highlighted. Parallel boxplots give univariate summary comparisons, while parallel coordinate plots potentially offer multivariate comparisons.

Figure 18 shows the default parallel coordinate plot of the ratios (without *Eq.TA* but with *logTA*), drawn for all the data except 55 outliers removed in an initial data cleaning. The companies that went bankrupt have been selected. The heavy overplotting may be obscuring information and the fact that every line is drawn the same whether it represents one case or many may also mislead. Nevertheless some features can be identified: all bankrupt companies had low values of *CL.TA*, high values of *Ebit.TA* and medium values of *Ebit.Int*; there are two unusual bankrupt companies (one a high outlier on *TL.TA* and the other a high outlier on *S.TA*).

One solution to get better discriminatory power is to use α-blending. A factor of 0.1 has been used in Figure 19, and it is now possible to see that the concentration of values for bankrupt companies on some of the variables applies to the bulk of the rest of the data as well, so that variables like *CL.TA* and *Ebit.TA* will not be as informative as might have been hoped. A final step can be to apply α-blending to the highlighting as well and that has been tried in Figure 20. This suggests that *Cash.TA* and *Intg.TA* might be more helpful than at first thought.

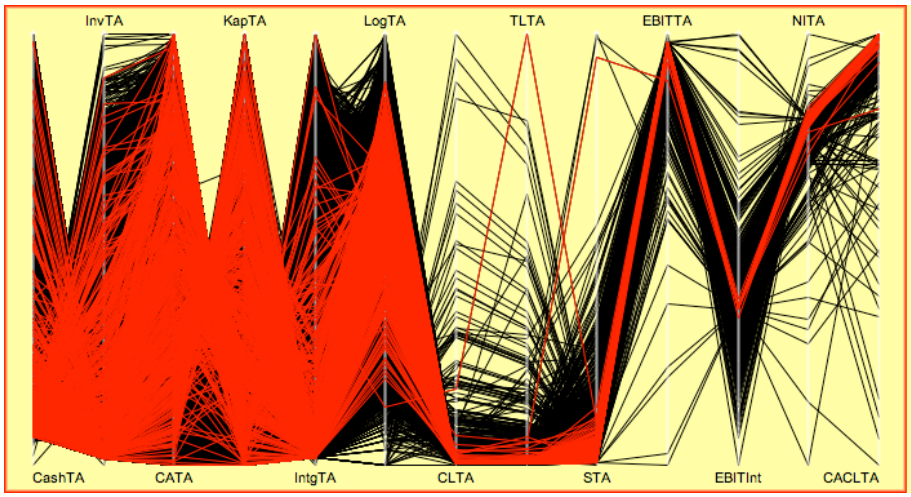

Fig. 18. Parallel coordinate plot of financial ratios and *logTA*, excluding 55 outliers, with bankrupt companies highlighted.

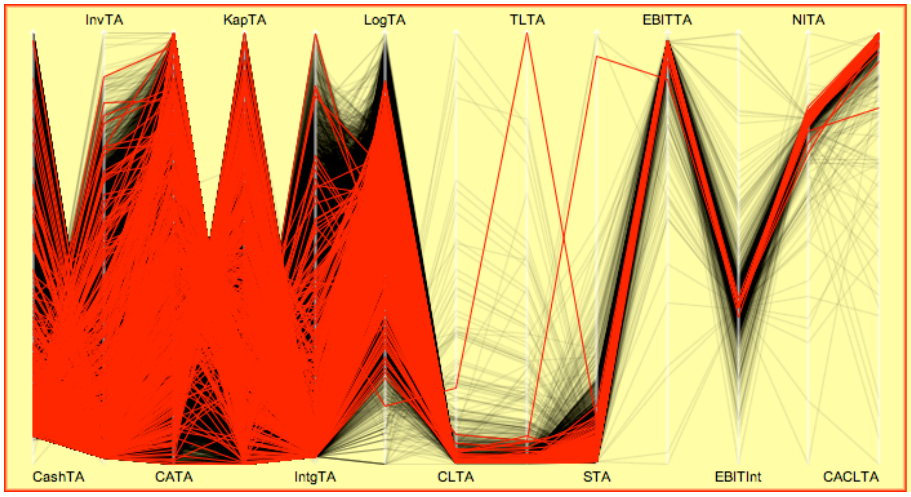

Fig. 19. Parallel coordinate plot of financial ratios and *logTA*, excluding 55 outliers, with bankrupt companies highlighted, α-blending=0.1 only for unselected data.

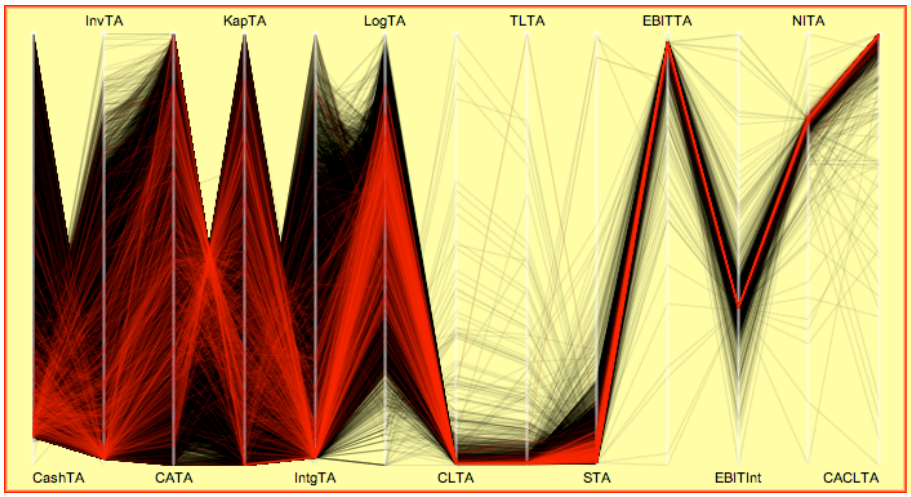

Fig. 20. Parallel coordinate plot of financial ratios and *logTA*, excluding 55 outliers, with bankrupt companies highlighted, α-blending=0.1 for selected and unselected data.

Using selection and linking for scatterplots can work well, but is dependent on the data distribution. Consider the two variables just mentioned above, *Cash.TA* and *Intg.TA*, and their distribution related to the companies which went bankrupt. A scatterplot could be drawn with a little α-blending, as on the left of Figure 21, or with a lot of α-blending as on the right. (α-blending has not been used for the highlighted cases, as on this scale they then disappear.)

Both plots contribute information but neither is fully satisfactory. Another alternative would be to draw a pair of scatterplots, one for each of the two groups of companies, but this is difficult to interpret. The success of particular plots depends on there being clear-cut information to find. The differences between the companies that went bankrupt and the others that did not are more complicated than can be displayed in a scatterplot of two variables.

Another alternative would be to use spinograms (as in Figure 24), but the numbers of bankruptcy companies are so small relative to the total number of companies that little can be seen. Fitted smooths would be better, though they require more computation.

The parallel coordinate plots employed here are a selection of many that might have been shown. The choice and order of variables influences what might be seen. The decision to discard some extreme outliers does too. The level of α-blending is also an influential factor. In other words, a parallel coordinate plot is like any other multivariate analysis, the user has a lot of freedom of choice. Careful thought helps, and so does statistical modelling. Having explored the data and built a model, parallel coordinate plots can again be useful, this time as a way to understand the model's relationship to the dataset.

8 Investigating Bigger Companies

Financial data for small companies is highly variable and could well be more unreliable than data for large companies, though this is difficult to assess. Certainly, large companies are different from small companies, and so studying

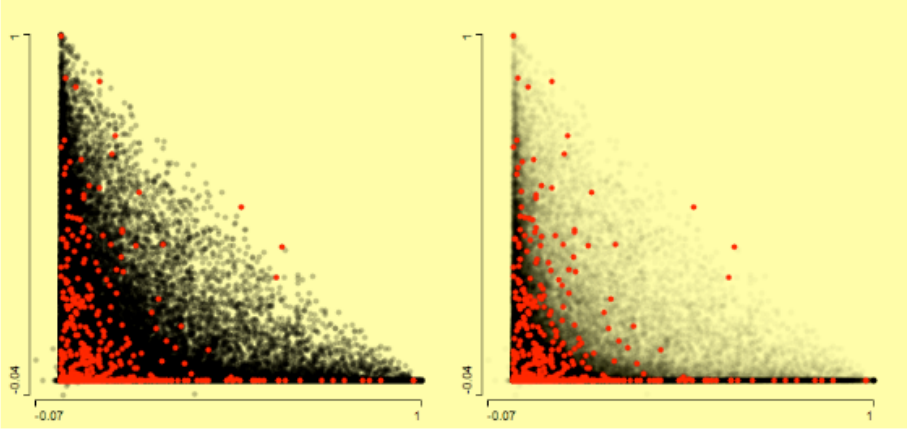

Fig. 21. Scatterplots of the ratios of intangibles and cash to total assets with companies that went bankrupt selected. More α-blending has been used in the right-hand plot.

them separately makes sense. In one important way they are not different: the bankruptcy rate for the biggest 7,690 companies (each with *Total assets* > 4000), 0.59%, is close to the rate for the rest of the dataset, 0.62% for the remaining 74,936 companies. On the other hand, using a looser definition of big (*Total assets* > 1000) gives bankruptcy rates of 0.73% for the 18610 'big' companies and 0.58% for the rest, a significant difference. Given the many different limits that might be used, a wide range of results is possible. However, it is not modest variations in bankruptcy rates by company size that are of interest, it is identifying which companies might go bankrupt.

Over forty years it would be reasonable to expect that the size of companies has increased. Curiously, for this dataset, the effect on *logTA* is negligible as Figure 22 shows.

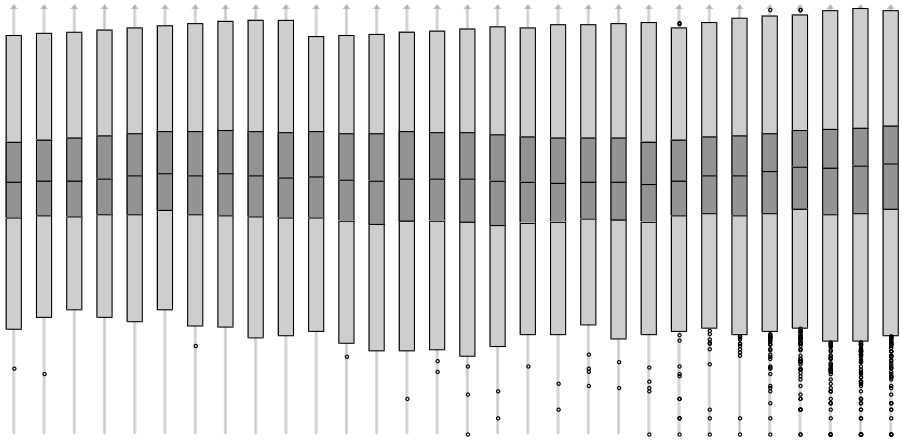

Fig. 22. Parallel boxplots of *logTA* by year from 1974 to 2003, all on the same scale.

Figure 23 shows boxplots of the original ratio variables for all the data with the group of big companies highlighted. Only the first five financial ratios are shown, as the distributions of the others are, as Figure 7 showed, far too skew to be informative. *logTA* is included to show the size distribution. Although the medians for the bigger companies differ noticeably from the median ratios for all companies (for cash, inventories and, hence, obviously, current assets, they are lower, and for the capital assets ratio the median ratio is higher), the distributions overlap substantially.

A more effective way of looking at the ratios is to use spinograms. In Figure 24 there are plots of the ratios *Cash.TA*, *Inv.TA*, *Kap.TA*, and *Intg.TA* with the big companies selected. The proportion of big companies declines as the cash ratio increases; it also declines as the inventory ratio increases, apart from the lowest group ($Inv.TA < 0.01$), where the proportion is relatively low; the proportion increases as the fixed capital ratio increases; the proportion is fairly constant for the intangibles ratio.

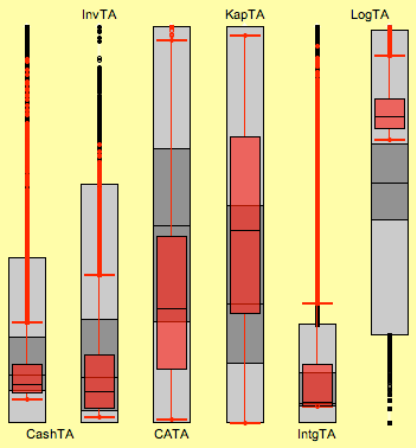

Fig. 23. Parallel boxplots of financial ratios and *logTA* for all companies. The background boxplots are for all the data and the superimposed standard boxplots are for the selected cases, companies with *Total Assets* > 1000.

Figure 25 shows just the data for the bigger companies for all variables with the bankrupt companies selected. Outliers are still affecting most of the second group of ratios, but two features stand out: *TL.TA* is generally higher, as we would expect from Table 7 for the whole dataset, and *Ebit.TA* is generally lower. It is tempting to draw conclusions about the fact that none of the biggest companies went bankrupt and that none of the companies that went bankrupt had a high cash ratio (even though the median is higher for these companies), but the selected cases make up such a small percentage of the total that caution should be exercised before drawing any conclusions.

9 Summary

Every data analysis is unique, because the data are always different. In the study reported here, there were mainly continuous variables (so parallel coordinate plots were useful); the few categorical variables had mostly large numbers of categories (so these had to be combined into groups); there was a fairly large number of cases (so approximating density estimations were helpful); and there were some variables that were highly skew (so that outliers and transformations were issues). A variety of plots were used, including barcharts, histograms, spinograms, boxplots, scatterplots, mosaic plots and parallel coordinate plots. Weighted versions of some plots also contributed. Trellis displays might have been tried, but then shingling of the conditional variables would have been required. That would be more appropriate after modelling. Interactivity, primarily selection, querying and linking, was used extensively, as is clear from the plots, but zooming and reformatting were also

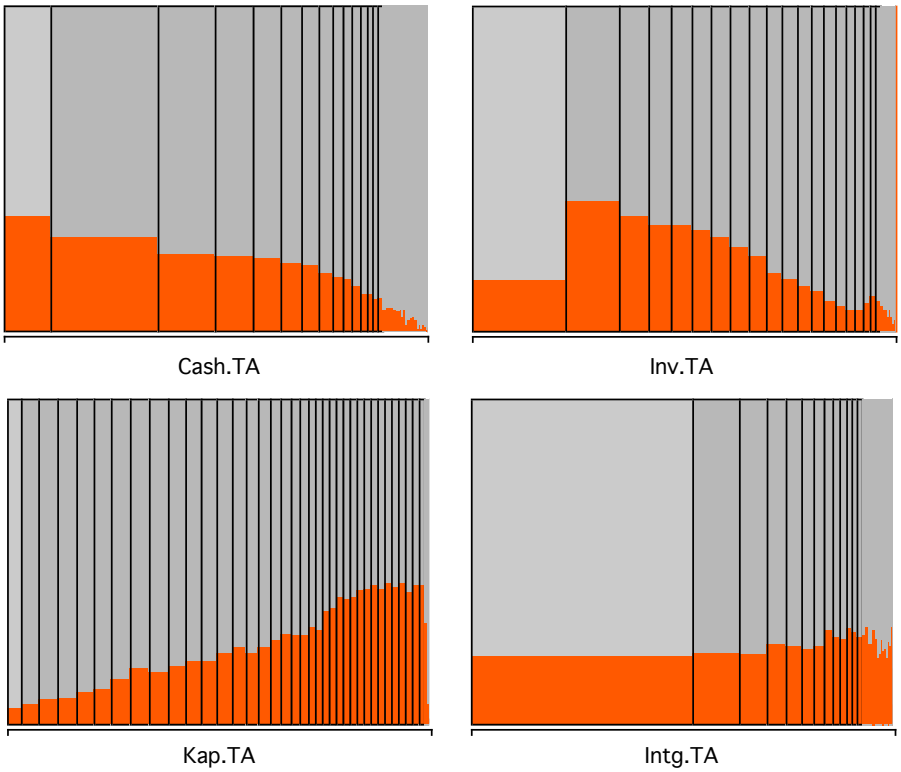

Fig. 24. Spinograms of the ratios *Cash.TA*, *Inv.TA*, *Kap.TA*, and *Intg.TA*. The companies with *Total Assets* > 1000 have been selected.

used a lot in the exploratory analyses. It is not easy to illustrate EDA in print and the chapter can only convey a pale shadow of the actual process.

Exploring the data is an important part of any data analysis. It is necessary to learn about the data, to check data quality and to carry out the data cleaning that is needed (and data cleaning is always needed with real datasets). EDA revealed here that there were some extreme outliers and some suspicious negative values. It underlined the need to transform some of the variables and it highlighted the geographic and sectoral structure of the dataset. It also revealed the surprising age of some of the data and the unexpected stability of the size distribution over time. Several interesting associations between variables were uncovered. Investigating the factors influencing bankruptcy gave further insights into the data and prepared the ground for statistical modelling.

Applying statistical models before exploring data is inefficient. Problems may arise because of some peculiarity of the data. Features are revealed that could have been found much more easily by just looking. The only possible

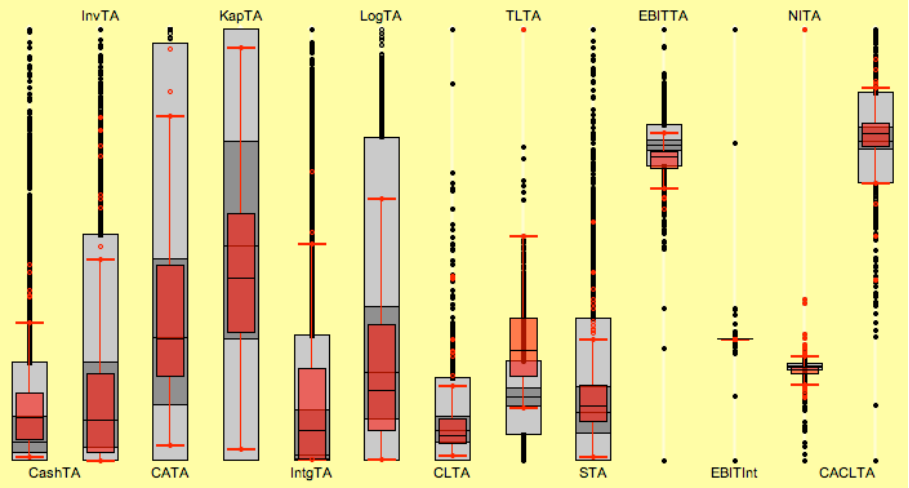

Fig. 25. Parallel boxplots of financial ratios and *logTA* for the 18610 companies with *Total Assets* > 1000. Companies that went bankrupt have been selected.

advantage of modelling without taking a look at the data first is the purist one of satisfying the idealist prerequisite for a hypothesis test — although whether that is relevant for any real analysis is moot.

Graphics are essential for exploratory work. They provide overviews and insights that complement statistical summaries. On the other hand, graphical analyses without follow-up analytic support remain inconclusive. Visualization results tend to be qualitative rather than quantitative, general rather than precise. Statistical modelling can assess the strength of evidence supporting ideas generated from graphical EDA and help to define those ideas more exactly. On top of that, statistical modelling can tease out more complex relationships, which are not immediately visually apparent. However, there is not much point in looking for complex relationships if the quality of the data is in doubt, and that is one of the reasons why modelling benefits from prior data visualization. Modelling also benefits from graphical support after analysis, both in investigating residual patterns for individual models and in comparing and combining groups of models. That is discussed in other chapters in the Handbook.

Software

The software used for most of the displays in this paper was Martin Theus's Mondrian (http://stats.math.uni-augsburg.de/Mondrian/). Other software was used at various stages to assist with data cleaning and restructuring.

Acknowledgements
Financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko" is gratefully acknowledged. Thanks also to Rouslan Moro, Dorothea Schäfer and Uwe Ziegenhagen for helpful comments on earlier drafts.

References

Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots, *Metrika* **51**(1): 11–26.

Hofmann, H. and Theus, M. (2006). Interactive graphics for visualizing conditional densities, *Journal of Computational and Graphical Statistics* .

Inselberg, A. (1999). Dont panic ... do it in parallel, *Computational Statistics* **14**(1): 53–77.

Playfair, W. (2005). *Playfair's Commercial and Political Atlas and Statistical Breviary*, Cambridge, London.

Tukey, J. (1977). *Exploratory Data Analysis*, Addison-Wesley, London.

Unwin, A. (1999). Requirements for interactive graphics software for exploratory data analysis, *Computational Statistics* **14**: 7–22.

Unwin, A. R., Theus, M. and Hofmann, H. (2006). *Graphics of Large Datasets*, Springer.